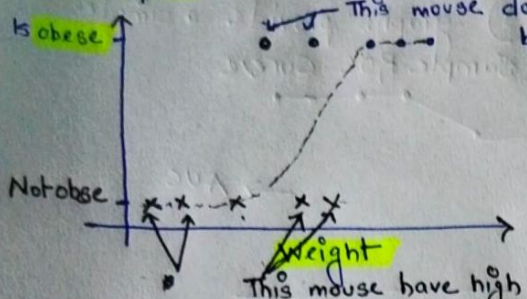
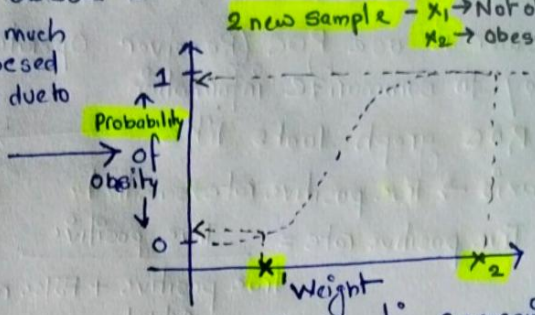
AUC and ROC

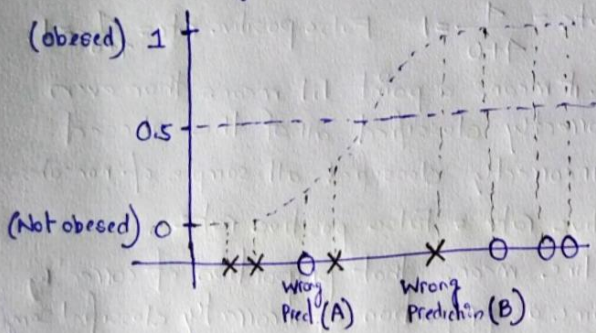
- Example - Whether the mouse is weight obese or not obese.





When we're doing logistic regression the y axis is converted to the probability of mouse is obese.

- However if we want to classify the mice as obese or not obese, then we need a way to turn the probabilities into classification.
  - One way to classify is to set a threshold at 0.5. probability > 0.5 then obese. probability <= 0.5 then not obese
- Suppose we get a four new set of mice who are NOT OBESE (x) and OBESE (0) respectively

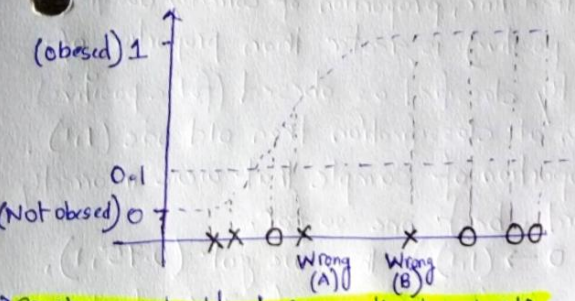


Predicted	Actual: Obesed	Actual: Not obesed
obesed	3	1
not obesed	1	3

Predict	Actual: Obesed	Actual: Not obesed
Obesed	TP	FP
Not obesed	FN	TN

- Suppose if is super important to correctly classify every obese sample, we can set the threshold to 0.1.

All 4 obesed (0) sample are now classified as obesed.

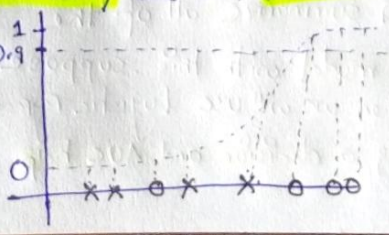


Predicted	Actual: Obesed	Actual: Not obesed
obesed	4	2
Not obesed	0	2

→ Lower threshold would also reduced the number of False - Negative (FN) because all the obesed mice are correctly classified. and it would reduce True Negative (TN) because two of the mice that were not obese were incorrectly classified as obese

→ So why we should decrease, the threshold? Suppose we are testing cancer or not cancer, then we should correctly classify who have cancer even if it means to decrease threshold.

- Otherway we can increase the threshold to 0.9 to correctly classify every non-obesed mice



Predicted	Actual: Obesed	Actual: Not obesed
Obesed	3	0
Not obesed	1	4

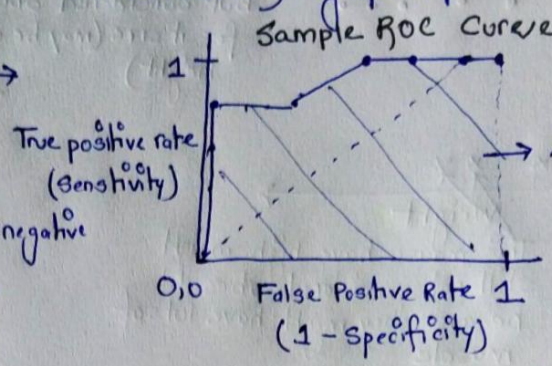
→ We don't have any false positive

→ With higher threshold, does a better job classify samples as obese and not obese.

- Threshold can be set anything from 0 to 1.
- How do we determine which threshold is best?

Answer is use ROC (Receiver Operator Characteristics) graphs provide a simple way to summarize information.

ROC graph looks like →



Yaxis → True positive rate → sensitivity

True positive rate = $\frac{\text{True positive}}{\text{True positive + False negative}}$

Predicted	Actual: Obesed	Actual: Not obesed
Is obesed	TP	FP
Not obesed	FN	TN

True positive rate tells what proportion of obese samples are correctly classified.

Xaxis → False positive rate = 1 - specificity = $\frac{\text{False positive}}{\text{False positive + True Negatives}}$

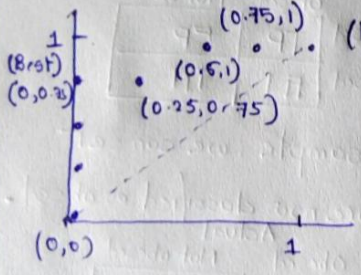
False positive rate tells proportion of non obese samples that were incorrectly classified and are false positive.

Suppose we set a threshold of 0.

Predict	Actual: Obesed	Actual: Not obesed
Obesed	4	4
Not obesed	0	0

True positive rate = $\frac{4}{4+0} = 1$ False positive rate = $\frac{4}{4+0} = 1$.

Point be (1,1). it means a point 1,1 means than even though we correctly classified all of the obesed samples, we incorrectly classified all samples of non-obesed.



→ Line shows True positive rate = False positive rate.
Any point on this line means that proportion of correctly classified obese is same as proportion of incorrectly classified samples that not obese.

→ Suppose a new point is calculated (0.75,1)
(0.75,1) is to left of dotted line, proportion of correctly classified samples were obesed (true positive) is greater than proportion of samples that are incorrectly classified as obesed (false positive)

In short, new threshold which lead to (0.75,1) is better at classification than old one (1,1).

→ New point (0.5,1), this further decreased the proportion of sample that were incorrectly classified as obesed (false positives). In otherword, best one so far.

→ As we are increasing the threshold (started with 0 → (1,1)) then we got (0.75,1), then (0.5,1) then (0.25,0.75), and (0,0.75). At (0,0.75) classify 75% of obese sample & 100% of the sample are not obesed. In other word, this threshold result in no false positives

→ The threshold is increased to 100% lead to (0,0) represent result is zero True positive & zero false positive (FP)

→ If we connect the dot, ROC graph is generated. ROC graph summarize (TP) all of the confusion matrices that each threshold produced.

→ Area Under Curve (AUC) is 0.9. so we can compare two model with this. suppose if AUC (logistic regression) is greater than AUC (random forest) then we will use logistic regression

→ ROC curve make it easy to identify best threshold for making a decision and AUC help which method is better.